DOCUMENT RESUME

ED 286 895 TM 870 503

AUTHOR Madhere, Serge
TITLE Evaluation Technique and Program Efficiency Measures: Statistical Derivations for the Regression Discontinuity Design.
PUB DATE Mar 86
NOTE 22p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Miami Beach, FL, March 12-15, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - General (140)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Compensatory Education; *Cutting Scores; Effect Size; Elementary Education; Evaluation Criteria; *Evaluation Methods; *Evaluation Utilization; Mathematical Models; Predictive Measurement; Pretests Posttests; *Program Effectiveness; Program Evaluation; *Quasiexperimental Design; Regression (Statistics); Remedial Programs; Statistical Inference; Statistical Studies
IDENTIFIERS Confidence Intervals (Statistics); *Efficiency Index; *Regression Discontinuity Model

ABSTRACT

One of the most appropriate quasiexperimental approaches to compensatory education is the regression-discontinuity design. However, it remains underutilized, in part because of the need to clarify the link between the mathematical model and administrative decision-making. This paper explains the derivation of a program efficiency index congruent with the regression-discontinuity design. The efficiency index is based on a confidence interval calculated for expected mean posttest performance. If the observed posttest mean corresponds exactly to the upper limit of the confidence interval, the efficiency index is +1 (optimal efficiency). If the observed posttest mean falls at the lower confidence boundary, the index value is zero (minimal efficiency). Data from a remedial mathematics program (grades 2, 3, 7, and 8) illustrate the possibilities for: (1) net growth (index above 1); (2) breakdown (index below 0); and (3) gradations of maintenance (index between 0 and 1). As conceived, the efficiency index is comparable to eta-square, the correlation ratio. Thus, its analytic context differs from that of the effect size coefficient commonly associated with classical control group design. A model is presented showing how variations in the size of the efficiency index may lead to different decision making options. (Author/LPG)

Evaluation Technique and Program Efficiency Measures:

Statistical Derivations for the Regression Discontinuity Design

Serge Madhere
Research Supervisor
Newark Board of Education

Presented at the Eastern Educational Research Association Meeting

Miami Beach, Florida
March 1986

Evaluation Technique and Program Efficiency Measures:
Statistical Derivations for the Regression Discontinuity Design

Abstract

One of the most appropriate quasi-experimental approach to compensatory education is the regression-discontinuity design. However, it remains under-utilized, and some suggest that its utility to program evaluation could be enhanced if the link was made more clearly between its mathematical rationale and the process of administrative decision-making (Linn, 1981). This paper explains the derivation of a program efficiency index congruent with the regression discontinuity design. As conceived, the efficiency index is comparable to eta-square, the correlation ratio. Thus, its analytic context differs from that of the effect size coefficient (Cohen, 1969) commonly associated with the classical control group design. We will further show how variations in the size of the efficiency index may lead to different decision-making options.

Evaluation Technique and Program Efficiency Measures: Statistical Derivations for the Regression Discontinuity Design

One of the most appropriate quasi-experimental approach to evaluate compensatory education is the regression-discontinuity design. However, it remains underutilized, and some suggest that its utility to program evaluation could be enhanced if the link was made more clearly between its mathematical rationale and the process of administrative decision-making (Linn, 1981). This paper explains the derivation of a program efficiency index congruent with the regression discontinuity design. As conceived, the efficiency index is comparable to eta-square, the correlation ratio. Thus, its analytic context differs from that of the effect size coefficient (Cohen, 1969) commonly associated with the classical control group design. We will further show how variations in the size of the efficiency index may lead to different decision-making options.

Perspective

Evaluation Design

The regression-discontinuity (Campbell and Stanley, 1966) is a quasi-experimental design appropriate for situations where there is a known interaction between treatment assignment and ability (achievement, aptitude, etc.). It has emerged in recent years as one of the most promising quantitative models for the evaluation of compensatory education. Based on the criterion of internal validity, the regression-discontinuity design has been shown to be superior to the norm-referenced model (Linn, 1981), since there often are multiple academic and contextual differences between the remedial group under study and the national sample from which test norms are developed. Based on the criterion of feasibility, the regression-discontinuity design has been found preferable to the classical experimental/control group approach, since it is impractical or unethical, in many instances, to withhold needed services from students in order to set up a comparison group (Wolf,

1981). Beyond the issue of applicability, the design may be most desirable, 1) when assignment to the 'treatment' group is based on a definite cutoff score, i. e., all students with a pretest score below a certain mark participate in the remedial program, while those above are dispensed of it; 2) when the educational environment includes multiple 'treatments,' and there is a need to separate the impact of the remedial, supplementary intervention from that of the general program of instruction. To determine the treatment's effectiveness, the task of the evaluator is to estimate what the performance level of the low achieving group would be without the remedial support, then, one tests to see whether the actual score for that group is significantly different from the expected value.

Two variants of this design exist. In the strict regression-discontinuity approach, separate pretest-posttest regression lines are obtained for the group above and the group below the cutoff point. Then two predicted values for that pretest cutoff score are calculated, by fitting it into each regression equation. A discontinuity in the regression lines, i.e., a difference between the predicted cutoff values, if significant, is taken as a measure of program impact. Tallmadge, Horst, and Wood (1975) propose a modification of the original technique that may be more sensitive to a possible pretest/program interaction among the low achieving students. In this version, known as regression-projection, the relationship between the pretest and the posttest is calculated only for the group of students above the cutoff score. Then, assuming linearity over the entire range of pretest scores, a single regression coefficient is used to estimate what the remedial group's posttest mean would have been under a 'no-treatment" condition. The formula for making such an estimate reads as:

$$E(\overline{Y}_t) = \overline{Y}_c + b_c (\overline{X}_t - \overline{X}_c)$$

[Insert Figure 1 here]

It simply means that the difference between the high achieving and the low achieving group on the posttest is expected to be the same as it was on the pretest, except for the imperfect correlation between the two measures. Any discrepancy between the projected and the observed posttest mean is attributed to the remedial treatment. The two versions of the regression design are illustrated in Figure 1. The details of the statistical test to establish significance of the differences can be found in Sween (1971) for the regression-discontinuity, and in Tallmadge and Horst (1976) for the regression-projection.

Statistical Analysis

The statistical tests offered to accompany the regression designs result in the usual t-value. But, as has been pointed out by many authors (Cohen, 1969; Hays, 1973), the emergence of a statistically significant value does not truly reveal the strength of the relationship between the independent and the dependent variable. The information provided by the index of significance is particularly limited when the hypothesis-testing paradigm is adopted. Hypothesis testing, however, is only one means of deriving statistical inference. As stated by Hays (1973), "in many circumstances," (and evaluation seems to be exactly one of these circumstances), "the primary purpose of data collection is not to test a hypothesis, but rather to obtain an estimate of some parameter" (p. 375). A range of values may be more useful or more stable than a single, unqualified estimate, given the presence of sampling error affecting

most research data. Rather than just ignoring the sampling error, an evaluator can place him/her self on safer ground by dealing straight forwardly with it, when drawing a conclusion about program effectiveness. To do that, one can turn to another form of statistical inference, the calculation of a confidence interval.

Ordinarily, in regression analysis, it is possible to establish confidence intervals for three different parameters: the regression coefficient itself, the actual score of an individual on the criterion measure, or the predicted value of a particular pretest score. Given the critical role accorded to the predicted mean value in the regression design, the calculation of the confidence interval is most necessary for that parameter. To obtain the boundaries of the confidence interval, i.e., the critical values for the expected mean, one can use the following formula adapted from Hays (1973):

$$\bar{Y}'_t = (t_{\alpha/2})\,(\text{est } \sigma_{yx})\sqrt{\frac{1}{N} + \frac{(\bar{X}t - \bar{X}c)^2}{NS_x^2}}$$

where: $\bar{Y}'_t$ = Predicted posttest mean for the treatment group

$\bar{X}_t$ = Mean of the treatment group on the pretest

$\bar{X}_c$ = Mean of the control group on the pretest

est σ_{yx} = The standard error of estimate adjusted by the sample size

$$\text{est } \sigma_{yx} = \sqrt{\frac{NS_y^2(1 - r^2)}{N - 2}}$$

For the t-value, any probability may be retained by the evaluator, depending on the desired level of confidence interval.

If the actual posttest mean for the treatment group does not fall within the calculated interval, one can be 95 percent confident that 'something extraordinary' is happening with the program. If the observed mean is above the upper limit of the confidence interval, the impact of the program is definitely positive. On the other hand, if the observed mean is below the

lower limit of the confidence interval, the return on the program is clearly not what one would expect. As one can see, the procedure is quite unequivocal about the extreme cases. One may say that it also increases the likelihood of arriving at a nonsignificant difference. But even within the region of nonsignificance, it is possible to set up a gradient of performance, which allows the evaluator to draw inferences not just about goal attainment, but also the level at which a program operates. Indeed, all the bits of information obtained from the standard statistical analysis can be condensed into one measure that we call the efficiency index. The term efficiency speaks of the average amount of progress made by the treatment group participants, relative to their own entry level and that of students in the control group. Mathematically, it is calculated according to the following formula:

$$E = \frac{(y - y')}{|y' - y''| + |y - y'|} + .5$$

where: $(y-y')$ = the difference between the observed (y) and the expected (y') posttest mean for the treatment group

$(y'-y'')$ = the difference between the expected mean (y') and its critical value (y'').

The absolute value of (y'-y") represents the distance from the lower limit of the confidence interval to its center, while the absolute value of (y-y') represents the distance from that center in either direction. Therefore, the first term in the mathematical expression simply defines the "gain" at posttest time corrected for uncertainty, i.e., the relative difference between the observed posttest score and the lower limit of the confidence interval; .5 is added simply to further facilitate interpretation.

Indeed, if the observed and the predicted posttest means coincide, the efficiency index will take the value of .5. If the observed posttest mean corresponds exactly to the upper limit of the confidence interval, the efficiency index will take the value of +1. If the observed posttest mean falls precisely at the lower boundary of the confidence interval, the efficiency index will take the value of 0.

Although the derivation of such an index may seem elaborate, its merit is that it tremendously simplifies the reporting of evaluation results to program administrators. That advantage can be appreciated when one has to deal with a program implemented at several grade levels. Whenever the efficiency index is greater than 1, the program is probably exemplary; whenever the efficiency index is negative, the program is probably in trouble. Even when the index falls between 0 and 1, (in other words, no statistical significance is obtained), it is still possible to call attention to different degrees of efficiency; in that sense, the procedure gets around the no-significant difference problem, the lack of sensitivity, that Stufflebeam et al. (1971) found as a frequent limitation of evaluation techniques.

The whole procedure is illustrated below with actual data obtained at four grade levels (2, 3, 7, and 8) for a remedial math program.

In grade 7, for example, students with a pretest score lower than 38 NCEs (29th percentile rank) were assigned to the remedial program. The average pretest score for this low achieving group was 30.64 NCE, compared to a mean of 57.49 for students not participating in the program. Based on the regression analysis, it was projected that the posttest performance for students in the first group would be around 25.8 NCE, in the absence of the remedial program.

$$Y'_t = 55.03 + .77\left(\frac{17.00}{12.01}\right)(30.64 - 57.49) = 25.78$$

A 95 percent confidence interval was calculated, that extends ± 7.04 NCE points around that central value.

$$25.78 \pm (2.001)\ (11.03)\sqrt{\frac{1}{59} + \frac{(30.64 - 57.49)^2}{59 \times (12.01)^2}} = 25.78 \pm 7.04$$

The observed posttest mean for the treatment group was 34.02, and fell outside of the confidence interval. It actually exceeded its upper limit by 1.20 NCE. That difference can be translated into an efficiency index equal:

$$E = \frac{34.02 - 25.78}{|7.04| + |34.02 - 25.78|} + .5 = 1.039$$

Clearly, the impact of the program is strongly positive at that grade level, for the average participating students.

[Insert Table 1 here]

The calculations for the other grade levels can be carried out in similar fashions.

Significance

To understand the utility of the efficiency index, we can show its relationship to other measures of treatment effectiveness, and to the administrative decision-making process.

A - Treatment Effectiveness

There exist several coefficients to indicate the impact of treatment on performance. They are mainly conceived in terms of the percentage of variance in the dependent variable accounted for by the treatment. In the framework of analysis of variance, when two conditions are involved and equal variances are assumed, the most appropriate indicator of impact may be omega-square

(Hays, 1973). In the framework of regression analysis, when a linear model may not entirely fit the data, the best suited measure of strength of association is eta-square, also called the correlation ratio. From these basic coefficients, one can derive other descriptive statistics that express the impact of a treatment in direct units of measurement rather than as a proportion. The effect size coefficient proposed by Cohen (1969) is such an indicator which clearly branches from omega-square. It expresses the difference between the means of a treatment and a control group in terms of the standard deviation (of the control group). The efficiency index proposed in this paper is more directly related to eta-square, the correlation ratio. Let's recall that:

$$\eta^2_{yx} = \frac{\sum_j n_j (M_{yj} - M_y)^2}{\sum_j \sum_i (Y_{ij} - M_y)^2}$$

where

the numerator stands for "the sum of squares between groups," and the denominator for the "total sum of squares." That denominator can be rewritten as:

$$\sum_j \sum_i \left[(Y_{ij} - M_{yj}) + (M_{yj} - M_y)\right]^2$$

The correlation ratio thus becomes:

$$\eta^2_{yx} = \frac{\sum_j n_j (M_{yj} - M_y)^2}{\sum_j \sum_i \left[(Y_{ij} - M_{yj}) + (M_{yj} - M_y)\right]^2}$$

Except for the summation signs and the power transformation, one can see that this mathematical expression is perfectly analogous to the ratio used in computing the efficiency index.

The preceding discussion already points to the differences between the efficiency index (EI) and the effect size (ES):

a) Computationally, their mathematical roots are distinct, with the former being linked to eta-square while the latter branches out from omega-square.

b) In terms of magnitude, the effect size coefficient expresses the distance between the mean of a treatment group and that of a control group, while the efficiency index measures the distance from the observed mean to the lower limit of the expected mean.

c) More importantly, the two measures belong to different contexts of analysis. There are some serious questions regarding the application of the effect size coefficient in situations calling for the regression-discontinuity design. Indeed, that design is equivalent to a repeated measure experimental condition, in which each subject receives the two available treatments (regular and remedial instructions). The two scores being compared (the predicted value and the observed value for the treatment group), cannot be considered entirely independent. To that extent, some limitations are placed on the anova framework and its associated statistics. The effect size coefficient, as we have seen, falls in that category. All this is to say that while the effect size maintains its legitimacy in the regular experimental-control group design, the efficiency index seems preferable with the regression-discontinuity design.

B - Management Information

Two questions need to be addressed now: 1) How does one convey that kind of complex information to administrators in a handy way? 2) How does one advance the probability that the reported information indeed be included in the decision-making process?

1 - Making it Accessible

Information on a program's efficiency may be reported in a modified scattergram as follows. The horizontal axis shows the pretest scores (say in NCE's) with a clear mark for the cutoff point; the vertical axis shows different values of the efficiency index. One can divide the area delineated by these axes into three subfields, by drawing two lines at point 1 and 0, perpendicular to the efficiency axis. The top line, at point 1, corresponds of course to the upper limit of the confidence interval calculated; it can be referred to as the optimal efficiency line. The bottom line, at point 0, corresponds to the lower limit of the confidence interval calculated; it may be referred to as the minimal efficiency line. The subfield above the optimal efficiency line is designated as a net growth area; the subfield between the optimal and the minimal efficiency lines is designated as a maintenance area;

the subfield below the minimal efficiency line is designated as a breakdown area. The points in the scatterplot represent the various sites or grade levels at which the program was implemented. If at a particular grade level the actual posttest mean falls within the confidence interval, for the predicted mean, that observation will appear between the two efficiency lines; this will suggest that the remedial program is operating as a maintenance unit, whose utility is to prevent the deterioration of skills, and thus sustain the operation of the regular instructional program; in other words, without it, the regular program of instruction may not be able to function with any kind of efficacy. If at another grade level the posttest mean exceeds the upper limit of the confidence interval, that observation will appear above the optimal efficiency line; this will suggest that the remedial program is operating as a production unit, capable of creating a net growth in students' competence. If at still another grade level the posttest mean fails to reach the lower limit of the confidence interval, that observation will appear below the minimal efficiency line; this will suggest that the remedial program is in disrepair. The whole procedure for reporting information on program efficiency is depicted in Figure 2.

2 - Making it Practical

In order to make the information he/she generates relevant to the decision-making process, the evaluator must have a good understanding of that process. That understanding should be based on empirical evidence about the overall program environment, and should also be be guided by a theoretical framework. Previous research (Baybrooke and Lindbloom, 1963) suggests that the process of rational decision-making follows four principles. What are those principles and what do they entail?

1. A decision requires a clear information base.

The information base, which is of course nothing other than previous evaluation results, may indicate one of three things: a) a given program is capable of producing net academic growth, i. e., its efficiency index is greater than 1; b) a given program operates as a maintenance unit, i. e., its efficiency index is between 0 and 1; c) a given program is experiencing a breakdown, i. e., its efficiency index is lower than 0.

2. A decision is always inscribed within a general approach to management.

Following Stufflebeam et al. (1971), we distinguish three possible approaches in an educational setting: a) a homeostatic approach, intended to sustain the achieved balance in a program; b) an incremental approach, aimed at "shifting the program to a new balance based upon small serial improvements" (p. 69); c) a neomobilistic approach geared for a large and significant change necessitated by critical program conditions.

3. A decision calls for selection or design of specific procedures to be followed.

This principle really speaks of the planning stage in the process. a) Planning may consist in simply standardizing or operationalizing the procedures presently in use. b) Another possibility is to target particular areas where the need is the greatest, or where resource allocation will be most efficient. c) Still another alternative is to reorganize a program in all its aspects, adjusting the objectives, providing new means, redefining personnel roles, setting check points for accountability.

4. A decision involves translating a set of selected procedures into activities in order to meet an objective.

Three courses of action may be followed: a) one can continue or recycle a set of practices proven to be successful; b) one can offer training and other activities in staff development; c) one can move to enforce or implement available guidelines/procedures where numerous discrepancies have been found between a program's objectives and modus operandi.

Stufflebeam et al. insist that the ultimate objective of a rational decision-making process, similar to the one outlined above, is educational improvement. While no educator would contest that view, it has been our experience that a number of immediate goals often supersede the ultimate objective. These immediate administrative goals fall into three categories: those aimed at producing change (transform-goals), those aimed at achieving control (conform-goals), those aimed at promoting or marketing a particular program or position for public relations purposes (inform-goals). These immediate goals, because of the rather quick payoffs associated with them, are the guiding lights of management. So, the evaluation results must be articulated to them in order to sensitize the decision-makers. We propose a restructuring of the decision-making model to reflect that situation. Figure 3 depicts this new structure.

The model establishes a correspondence between each immediate goal and the type of elements in the decision-making process which it seems most congruent with. It can be of great utility to the evaluator in formulating his/her recommendations for program development. Depending on the kind of evaluation results obtained (i. e., the value of the efficiency index), a particular administrative approach, some specific planning procedures, and a set of corrective/supportive activities may be suggested. That kind of

detailed, facilitative work has a good probability of catching the attention of the decision-makers.

References

Baybrooke, D. and Lindbloom, C. E. A Strategy of Decision. New York: Free Press, 1963.

Campbell, D. T. and Stanley, J. C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences. Academic Press, New York, 1969.

Hays, W. L. Statistics for Psychologists (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Linn, R. L. Measuring Pretest-Posttest Performance Changes. In R. A. Berk (ed.): Educational Evaluation Methodology. Baltimore: John Hopkins University Press, 1981.

Stanley, J. C. Reliability. In R. L. Thorndike (ed.): Educational Measurement. American Council on Education, Washington, D.C. 1971.

Stufflebeam, D., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond. R. L., Merriman, H. O., and Provus, M. M. Educational Evaluation and Decision Making. Peacock Publishers, Itasca, Illinois, 1971.

Sween, J. A. Experimental Regression Design: Inquiry into the feasibility of nonrandom treatment allocation. Unpublished doctoral dissertation, Northwestern University, 1971.

Tallmadge, G. K., Horst, D. P., and Wood, C. T. Practical Guide for Measuring Project Impact on Student Achievement. Government Printing Office, Washington, D. C. 1975.

Tallmadge, G. K., and Horst, D. P. Procedural Guide for Validating Achievement Gains in Educational Projects. US HEW, Washington, D. C., 1976.

Wolf, R. M. Selecting Appropriate Statistical Methods. In R. A. Berk (ed.) Educational Evaluation Methodology. Baltimore: John Hopkins University Press, 1981.

Table 1

Statistical Data for Chapter I and Nonchapter I Students in Mathematics

Parameters	Grade 2		3		7		8	
	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.	Treat.	Cont.
1. Pretest Mean	32.04	64.80	23.27	60.00	30.64	57.49	30.09	56.83
2. SD of Pretest	11.26	14.89	9.91	16.73	8.31	12.01	9.36	14.46
3. Posttest Mean	37.70	58.94	32.98	59.13	34.02	55.03	37.40	56.02
4. SD for Posttest	17.27	19.39	10.95	16.31	11.88	17.00	8.15	14.58
5. Cutoff Score	41.90	-	28.20	-	38.00	-	38.00	-
6. Pre-Post Correlation	-	.57	-	.39	-	.77	-	.59
7. Sample Size (N)	70	65	64	61	58	59	66	60
8. Expected Post Mean	34.75		44.70		25.78		40.12	
9. Confidence Interval for (8)	±9.62		±10.20		±7.04		±6.67	
10. Efficiency Index	+.734		-.034		+1.04		+.21	

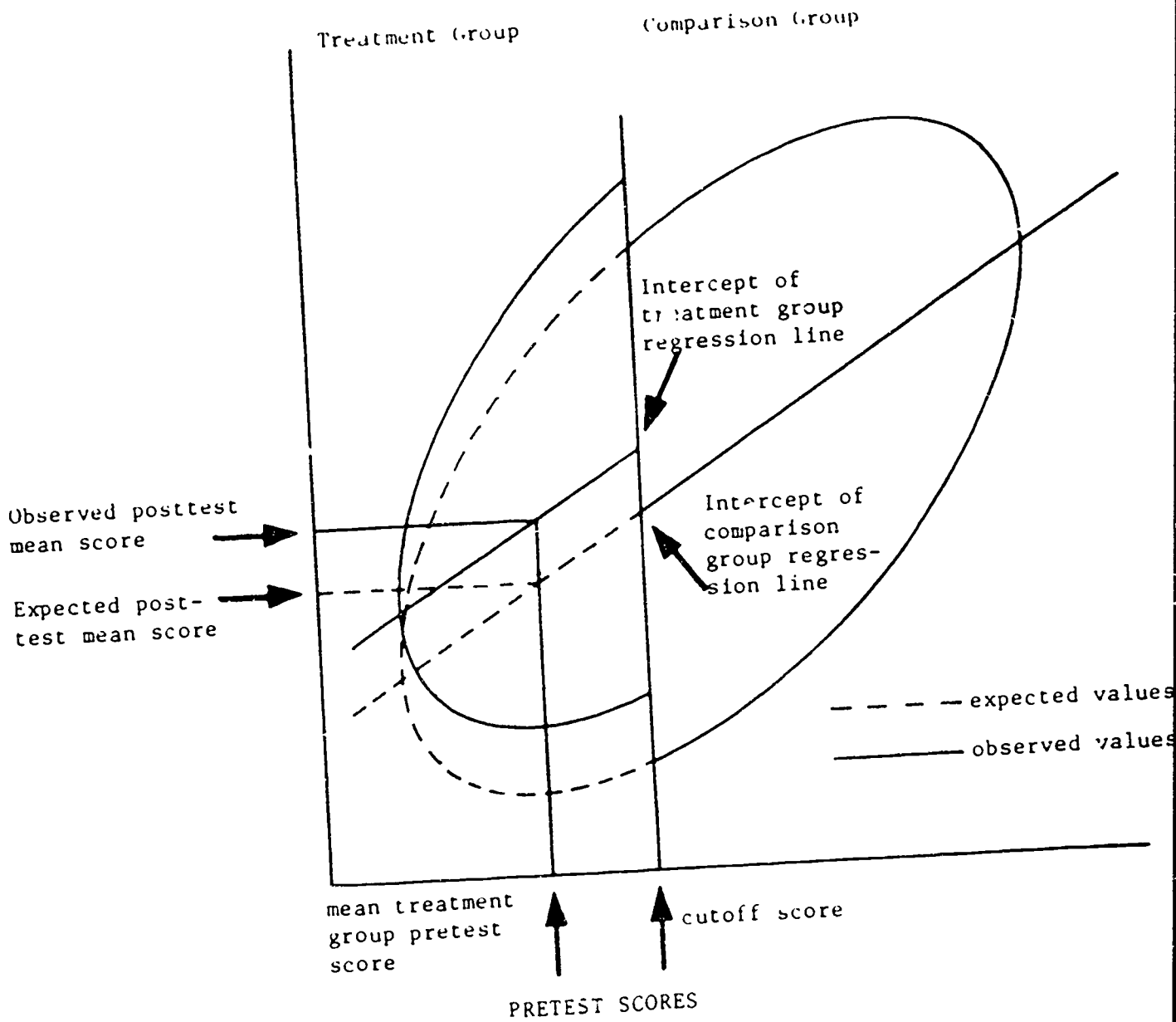

Fig. 1. Score distributions with treatment effect independent of pretest status. (reprinted from Tallmadge and Horst, 1976)

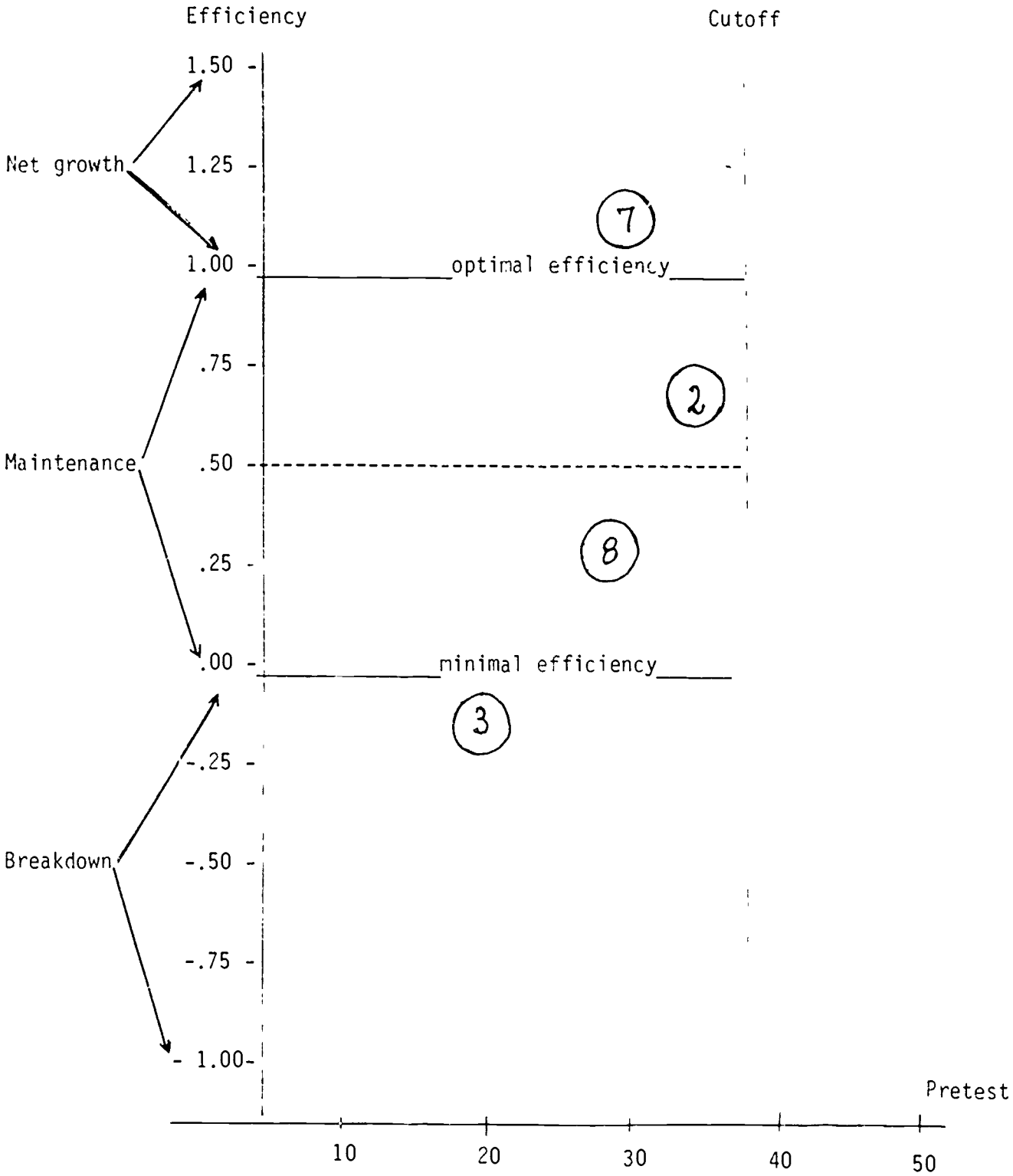

FIGURE 2 - PROGRAM EFFICIENCY AT FOUR GRADE LEVELS

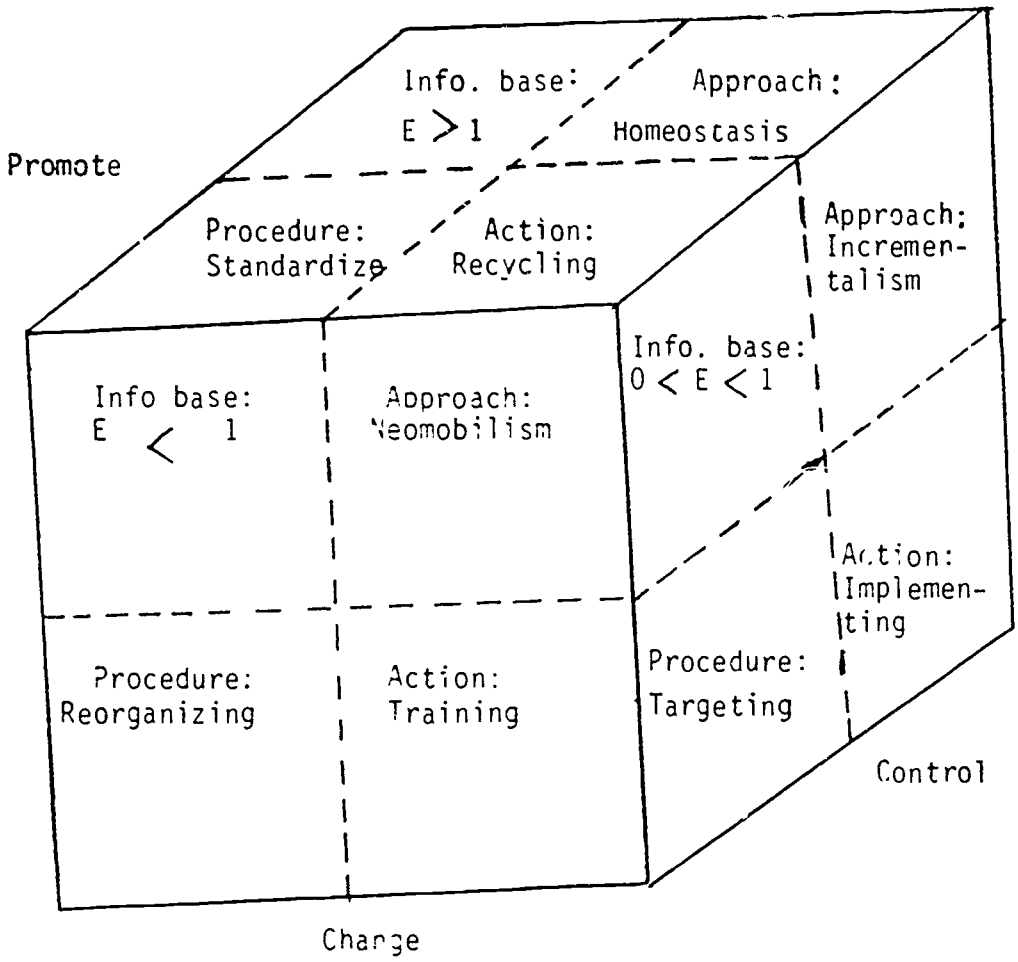

FIGURE 3 - AN EVALUATION-BASED MODEL FOR DECISION MAKING